DISAMBIGUATING NEAR SYNONYMS IN MEDICAL DISCOURSE
A multilayered corpus analysis of *disease*, *illness* and *sickness* in the British National Corpus

STEFANIA M. MACI, RÉKA R. JABLONKAI, MAREK ŁUKASIK, SOPHIKO DARASELIA, DANIEL KNUCHEL[‡]

Abstract - This paper discusses the preliminary results of a corpus-based analysis of three basic health-related lexical items: *disease*, *illness* and *sickness* on the British National Corpus (henceforth BNC) CQP Web platform (2007 XML). Synonymous at first glance, the terms exhibit a certain degree of co-text and context semantic variation; therefore, the lexical items in question cannot be used interchangeably. This in turn may pose some difficulties in inter-lingual translation and language learning, mainly stemming from the lack of full equivalence (or, in some instances, zero equivalence) between the words and their counterparts in some other languages, such as German or Italian. The paper aims to demonstrate how collocational behaviour and semantic profiles can help disambiguate near synonyms along a cline between general and specialised discourse. To this end, the study employed corpus linguistic methods and analysed the BNC across all its text genres. The collocational patterns of the three selected lexical items were examined in the corpus and the semantic profiles of the lexical items were established. The findings suggest that the three health-related near synonyms exhibit markedly different collocational behaviours and semantic preferences. It is therefore suggested that the approach adopted in this study could be applied to help disambiguate the meanings of near synonyms appearing in any specialised discourse at both intra- and inter-linguistic levels. Future research will compare the findings resulting from a similar investigation to be carried out on COCA to see the extent to which, if any, (a) meanings can vary and (b) whether meaning variations associated with these items depend on the interactants (i.e. professionals/laymen).

Keywords: medical discourse; corpus-based analysis; statistics; semantic preference; semantic profile; collocation.

[‡] The research is the result of a joint effort of the five authors. However, Daniel Knuchel and Marek Łukasik wrote the *Introduction*, *Method* and *Conclusions* sections; Réka Jablonkai wrote the *disease* paragraph the *Methods* and *Conclusions* sections; Stefania M. Maci revised the *Introduction*, *Methods*, *Conclusions* sections, and wrote the *sickness* paragraph and the *Conclusions* section; Sophiko Daraselia wrote the *Illness* paragraph and the *Conclusions* section. The investigation carried out by Stefania M. Maci is part of the MIUR-funded research project of national interest (PRIN) *Knowledge Dissemination across media in English: continuity and change in discourse strategies, ideologies, and epistemology* (COFIN grant no. Prot. 2015TJ8ZAS_002), directed by prof. Marina Bondi of the University of Modena and Reggio Emilia – local coordinator, prof. Maurizio Gotti.

Doi Code: 9788883051531p127

1. Introduction

The history of English medical discourse dates back to the 17th century, when the role of Latin diminished in favour of vernacular languages. This coincides with some important changes in the medical profession at the time, mainly the shift from a scholastic (authority and/or dogma-based) to an evidence-based approach (Taavitsainen 2018). Indeed, as Taavitsainen (2018, p. 252) explains, while genre conventions change slowly, some elements acquire new connotations. Evidence-based medicine and its approach mean the practice of medicine is hierarchically based on scientific evidence, with a crucial role attributed to "patient values and preferences in clinical decision making, and the development of the methodology for generating trustworthy recommendations" (Djulbegovic and Guyatt 2017, p. 415). In this context, there is a need for reliable and correct information, because humans 'consume' information, or better, "[h]umans are "informavores" - we need evidence to effectively function in the world around us" (Djulbegovic and Guyatt 2017, p. 421).

Also, new discoveries, novel procedures, methodologies and sophisticated equipment have led to an exponential increase in the numbers of terms and new medical genres, all of which have been necessary for the precision-oriented communication of professional knowledge (Gotti 2016).

At the beginning of the 20th century, the language of medical science around the world still used three languages to roughly equal extents: German, English and French (Baethge 2008). However, similar to many other specialised fields (Crystal 2002), English took the lead in the mid-20th century, particularly in scientific publications, and its dominance has prevailed ever since in worldwide professional medical communication (Baethge 2008). On the other hand, national languages are still widely used locally, in doctor-patient communication in individual countries, in teaching and in scientific activity; furthermore, the dominance of English on the Internet in non-English speaking countries has declined, as evidenced by Baethge (2008) and Graddol (2006, p. 14). The use of English in scientific publications and native languages on the Internet corroborates the presence of some form of linguo-pragmatic dichotomy, which is rooted in both systematic and linguistic differences (that still persist). For example, anatomical terms and the names of diseases are imported directly with their correct Latin endings into Germanic languages, such as German or Dutch, while they are more readily naturalised in Romance languages, such as French or Italian (Wulff 2004). In addition, languages can also borrow medical terms from English. Borrowings also occur between languages other than English (Wulff 2004). These processes result in a number of inconsistencies (at various levels) and paint a highly complicated picture of

current medical language, which might prove challenging for professional medical translators. At the same time, health communication contexts in English-speaking countries are becoming increasingly multilingual while English is often used as a lingua franca in doctor-patient communication or between medical professionals (Sentel and Braun 2012). As many doctors, nurses and other healthcare workers are non-native speakers of English, special attention should be paid to their use of English in healthcare contexts and it should be ensured that they use the right expressions and terms when communicating (Candlin and Candlin 2003). It is vital that misunderstandings or imprecise uses of language are avoided as these might complicate diagnoses and/or disease treatment.

The aim of this study is to explore how corpus linguistic methods can be applied to disambiguate the meaning of the selected health terms. This study is part of a larger-scale project that investigates the lexical behaviour and semantic profiles of some selected health terms in several other languages which belong to the same or different language families. We hope to gain insights into potential translation problems of medical terms and phrases from English into other languages, for instance, Georgian, German, Italian, Hungarian and Polish, and vice versa.

More specifically, this study aims to identify the core meanings of three near synonymous lexical items in English, namely: *disease*, *illness* and *sickness*. These basic health-related words are investigated through their collocations generated for the entire British National Corpus (henceforth BNC) CQP Web platform (2007 XML). Collocations constitute the basis on which meaning analyses can be carried out, given that they highlight the most frequent semantic fields within which each lexical item can be grouped.

2. Near synonyms and semantic preference

Our study is based on the assumption that meaning is a pragma-semantic construct (Wittgenstein 2003; Busse 2015, pp. 91-122). Therefore, although words may seem to be synonymous, they might differ in their use. The relationship between such words is often referred to as near synonymy in the literature. Xiao and McEnery (2006) define near synonyms as "lexical pairs that have very similar cognitive or denotational meanings, but which may differ in collocational or prosodic behaviour" (p. 108). Moreover, such near-synonymous words have specific semantic profiles which we understand as cognitive and denotational meanings, plus their use in context (which grants the speaker pragmatic knowledge). In other words, the semantic profile of a word is a broad meaning-driven sketch of this word. Especially in language learning, it is crucial for non-native speakers to have pragmatic knowledge of

L2 in general and to know the word sketches for near synonyms (Barron 2003, cf. also Baker *et al.* 2013). In particular, when there are no equivalents in L1 and L2, near synonyms are difficult to translate. An example is the German word *Krankheit*: although it can be translated into English as *disease*, *illness* or *sickness*, it is fundamental to distinguish their different semantic profiles, which will help users to use the words appropriately and express intended meanings precisely. Other examples include Italian words, such as *malattia*, which can be translated into English as *disease*, *illness* or *sickness*. As aptly underlined by Loiacono (2018, p. 398), the distinction between the terms *illness* and *disease* is an endogenous/ exogenous one, so that the term *illness* should refer to the state or condition of the disease, whereas the term *disease* refers to the type of disease itself. As a consequence, this suggests differences in frequencies of the singular/ plural forms of the two items. Yet, there are blurred cases due to external social forces, especially among laymen. The issue is not simply a linguistic problem: things are far more complex, because the distinction between *illnesses* and *diseases* is the distinction between patients and professionals:

> Patients suffer "illnesses", doctors diagnose and treat "diseases" […] when physicians dismiss illness because ascertainable "disease" is absent, they fail to meet their socially assigned responsibility. It is essential to reintegrate "scientific" and "social" concepts of disease and illness as a basis for a functional system of medical research and care. (Engberg 1977 in Loiacono 2018, p. 399)

A clear understanding of the semantic profiles of the English terms will facilitate the selection of the most appropriate equivalent in any given context.

To understand the semantic profile of a word we can look at "the company it keeps" (Firth 1957 [1951], p. 11). Firth introduced the term collocation to describe word associations and their impact on meaning. In the last six decades, several studies on collocation have been published and different empirical methods have been tested for the analysis of collocational patterns of lexical items (cf. Brezina *et al.* 2015; Gablasova *et al.* 2017). In this study, we align with Sinclair's (2004, p. 28) definition of a collocation as "a frequent co-occurrence of words".

Previous analyses of collocates of particular lexical items have found that lexical items have a tendency to co-occur with "other words that belong to a particular semantic set" (Hunston 1995, p. 137). Stubbs (2001), for example, defines semantic preference as "the relation, not between individual words, but between a lemma or word-form and a set of semantically related words" (p. 65). To illustrate this category of relation, Partington (2004) gives the semantic preferences of *sheer*. The semantic sets the word *sheer* was found to collocate with included (1) "magnitude", "weight" or "volume", e.g.

the sheer volume of reliable information; (2) "force", "strength" or "energy", e.g. *the sheer force of his presence*; (3) "persistence", e.g. *sometimes through sheer insistence*; (4) "strong emotion", e.g. the sheer joy of life and (5) physical quality, e.g. *the sheer glamour of evil* (p. 145). Furthermore, he demonstrated that there is also interaction between typical syntactic behaviours of words and their semantic preferences. In the example of *sheer*, the typical structure for the first two semantic sets, that is, "magnitude" and "force" words, was found to be "*the sheer* (noun phrase) *of* (noun phrase)". In the third semantic category, the word *sheer* was found to be often preceded by prepositions expressing means or manner, e.g. *through, out of, by*. Nelson (2006) found that, in business discourse for example, the word *package* had a preference for being connected to computers, and it also shared a preference related to finance, with words like *merger* and *market.*

Stubbs (1995) also demonstrates that lexical items have a tendency to co-occur with negative or positive words. This phenomenon is usually referred to as semantic prosody in the literature (Partington 2004). In his analysis of the word *cause*, for example, Stubbs (1995) found that its most frequent collocations are negative abstract nouns like *anxiety, concern* and *crisis,* and many examples are from the medical field, like *cancer, blood, death,* and *disease*. Furthermore, Nelson (2006) claims that looking at the semantic prosody of words as used in business discourse not only reveals insights into language use, but also provides information about the business world as such. The examples he gives here are semantic prosodies of the words *boss* and *manager*. According to his findings, *boss* has a tendency to be used with negative adjectives, such as *meanest* and *old-fashioned*, whereas *manager* displays positive collocates, like *excellent* and *good* (Nelson, 2006). In addition, Partington (2004) also analyses the relationship between semantic preference and semantic prosody. He suggests that in most cases semantic prosody can be considered a sub-category of semantic preference, a special case that includes "instances where a lexical item shows preference to co-occur with items that can be described as bad, unfavourable or unpleasant, or as good, favourable or pleasant" (p. 149). In a further analysis, however, he notes that "semantic preference is a 'narrower' phenomenon - relating the node to another item for a particular semantic set - than prosody which can affect wider stretches of text" (p. 151). He also illustrates how semantic preference contributes to building semantic prosody and how prosody in turn restricts the preferential choices of the node word. Several authors (Baker *et al*. 2008; Bednarek 2008), however, argue that a clear distinction should be made between semantic preference and semantic prosody. Bednarek (2008) proposes that, following Sinclair (2004), the term semantic preference should exclusively be applied to "collocations of a lexical item with (more or less specific) semantic subsets" (p. 132), and the term semantic prosody used for

all other attitudinal and evaluative meanings which often go beyond being merely positive and negative. The present study focuses on semantic preference when attempting to identify nuanced differences in meaning and usage between the selected health-related lexical items.

3. Methodology

3.1. Selection of lexical items

The terms selected for the analysis in the current study were chosen on the basis of two independent pilot analyses: (1) a comparison of dictionary definitions of the health-related lexical items, (2) synonym-finding query applied to the BNC (BYU-BNC at corpus.byu.edu, see Davies 2004).

3.1.1 Dictionary definitions

Definitions of the terms *disease*, *illness* and *sickness* were analysed as regards their synonyms in three online English dictionaries, namely, the Oxford dictionaries, including the Oxford English Dictionary, the Collins English Dictionary, and one of the most popular online medical dictionaries, i.e. the medical Merriam-Webster Dictionary (https://www.merriam-webster.com/medical). Although the investigation is based on the BNC, the fact that the medical Merriam-Webster Dictionary is American does not pose any problem. Indeed, the last attested time when the medical community stressed the linguistic importance of any difference existing between the British Medical Dictionary and the Webster American Dictionary was in 1962 (Talbott 1962), and ever since it has not been dealt with.

The results indicate that, overall, the selected lexical items are perceived in general language as being synonymous in relation to one another, with the reservations that (a) the study should be treated as approximate, insofar as the strength of synonymy relations is not provided in any of the aforementioned dictionaries and (b) in some cases one synonym is simultaneously offered as the *genus proximum* in the definition (i.e. in the *definiens* part), which results in a circular definition rather than an indication of a semantic position of the *genus* against the *definiendum*.

The term ‘disease’ was chosen as the point of departure, being the most generic and overarching medical term to represent the concept of interest, namely, that of “a disorder of structure or function in a human, animal, or plant, especially one that produces specific symptoms or that affects a specific location and is not simply a direct result of physical injury” (cf. https://en.oxforddictionaries.com/). Oxforddictionaries.com presents the

terms *illness* and *sick* as synonyms of *disease*. An 'illness' is "a disease or period of sickness affecting the body or mind"; sickness is "[t]he state of being ill" (https://en.oxforddictionaries.com/). In *Collins*, a 'disease' is an "illness which affects people, animals, or plants, for example one which is caused by bacteria or infection". An 'illness' is a particular disease, such as measles or pneumonia. *Sickness* is the state of being ill or unhealthy (https://www.collinsdictionary.com/).

The Merriam-Webster medical dictionary gives the following definition for 'disease': "an impairment of the normal state of the living animal or plant body or one of its parts that interrupts or modifies the performance of the vital functions, is typically manifested by distinguishing signs and symptoms, and is a response to environmental factors (as malnutrition, industrial hazards, or climate), to specific infective agents (such as worms, bacteria, or viruses), to inherent defects of the organism (as genetic anomalies), or to combinations of these factors: sickness, illness". The condition of having a *disease*, therefore, is that of having a combination of the two factors of sickness and illness. An 'illness' is defined as the "unhealthy condition of body or mind: sickness and 'sickness' as "the condition of being ill: ill health" or the condition of having a "specific disease" (https://www.merriam-webster.com/medical).

According to the OED, these words entered the English language and became part of the English lexicon at different times. In particular, the term 'sick' is a common Germanic word and is attested to in Old English (from 700 AD), whereas 'ill' is used in Early Middle English (from 1200 AD) and 'disease' (from 1300 AD). The words have developed senses that were associated with some of their first meanings and usage. For example, *ill* has been synonymous with *evil* (although not etymologically related) from the 12th century, which resulted in different meanings from 'sick' or 'disease'. For the purposes of this research, we will compare OED senses with those of the BNC corpus. In this analysis, we do not consider obsolete meanings as these do not add much information to this comparative study.

The word 'disease' originated in Middle English (1150 to 1500), meaning the "absence of ease, uneasiness, discomfort" (OED). In the OED, *disease* as a noun has three distinctive meanings:

1. Absence of ease; uneasiness, discomfort;
2. A condition of the body, or of some part or organ of the body, in which its functions are disturbed or deranged; Also applied to a disordered condition in plants;
    a. The condition of being out of health; illness, sickness;
    b. An individual case or instance of such a condition; an illness, ailment, malady, disorder;

   c. Any one of the various kinds of such conditions; a species of disorder or ailment, exhibiting special symptoms or affecting a special organ.
3. *Figurative*: A deranged, depraved, or morbid condition (of mind or disposition, of the affairs of a community, etc.; an evil affection or tendency.

The term *illness* derives from the adjective *ill,* introduced during the Middle English period with the sense of 'morally wicked'. Its meaning as a substantive refers to the quality or condition of being ill (in various senses). The OED records it with the meaning of "bad or unhealthy condition of the body (or, formerly, of some part of it); the condition of being ill; disease, ailment, sickness, malady".

The term 'sickness' derives from the adjective 'sick', with the sense of 'suffering from physical ailment', and was introduced during the Old English period. The substantive has four distinctive senses indicated by the OED as follows:

1. The state of being sick or ill; the condition of suffering from some malady; illness, ill-health (also figuratively);
2. A particular disease or malady (also in a figurative sense). It may also refer to a defect in wines or to a disease in sheep which can cause braxy;
3. A disturbance of the stomach manifesting itself in retching and vomiting;
4. Figuratively, it indicates utter disgust or weariness.

Although all dictionaries indicate that a *sickness* is a state or a temporary condition, they also suggest that the terms are not interchangeable. Yet they have not been useful in defining these words: *oxforddictionaries,* for instance, has indicated that *disease* is the overarching term, and that *illness* can have as synonyms both *disease* and *sickness,* while *sickness* can be a synonym only of *illness*. The OED reveals that *disease* has as synonyms both *illness* and *sickness,* and that synonyms of *illness* can be *disease* and *sickness*, but the synonym of *sickness* can only be *illness. Collins,* on the other hand, suggests that *disease* is an *illness* and vice versa, but that the condition of *sickness* is given by an *ill* state. The medical Webster-Merriam Dictionary indicates that a *disease* can be a *sickness* or an *illness*, that an *illness* is a *sickness* and that a *sickness* is a condition reflecting a *disease* in which one person is *ill*. Clearly, the use of dictionaries is not enough and this must be implemented with synonym-finding queries in the BNC.

3.1.2 Synonym-finding queries in the BNC

Synonyms were also found in the BNC via synonym-finding queries carried out through the corpus.byu.edu platform. The lemma rather than the word-form was chosen as a query node. The corresponding absolute frequencies of the three lemmas are as follows: DISEASE (f=8,799 singular; 1,817 plural), ILLNESS (f=3,194 singular; 506 plural) and SICKNESS (f=1,186 singular; 14 plural).

In a search of the whole corpus for synonyms of DISEASE, *illness* ranks first, with *sickness* coming eighth on the rank list, preceded by *disorder* (f=1,604), virus (f=1,474) and syndrome (f=1,197). In an analogous query for ILLNESS, *disease* comes first, with *sickness* ranking tenth, preceded, again, by *virus* and *syndrome*. The BNC (BYU version), apparently, does not help us to disambiguate these terms. For this reason, a more in-depth analysis is necessary.

3.2. The corpus

The research is based on queries applied to the entire British National Corpus through the University of Lancaster UCREL CQP Web platform[1] (see Hardie 2012). CQP Web offers access to the 2007 XML edition of the BNC, which comprises 112,102,325 word tokens and 638,862 word types, derived from 4,048 text samples. The corpus is POS-tagged using the BNC Basic Tagset (also known as C5),[2] and offers rich metadata, allowing the researcher to compare fine-grained sets of data across various categories (parameters). The simple search mode allows queries of the entire corpus or the written/ spoken mode only. The system also supports more advanced searches using Simple Query Syntax.

3.3. Methodological approach

With the aim of drafting semantic profiles of the three lexical items under investigation, we conducted a series of lexical analyses, mainly with the use of simple frequency counts (absolute and relative frequencies, AF and RF, respectively), dispersion measures (mean frequency, standard deviation and Juilland's D (distribution tests) and collocation extraction statistics (log likelihood and log ratio), as described below.

[1] https://cqpweb.lancs.ac.uk/ (17.3.2018).
[2] http://www.natcorp.ox.ac.uk/docs/c5spec.html (24.3.2018).

Dispersion measures are necessary in order to avoid biased results for AF and RF. Indeed, they helped us to compute frequencies of occurrence and co-occurrence of the three lemmas under investigation. Since in isolation statistical tests may be misleading, as each single test does not take into consideration the degree of dispersion, the three tests were used together to yield sound results (cf. also Gries 2008).

Collocates are words which usually go with the word under investigation (the node) and are computed within a range distance from the node of 5-1 words to the left and right of the node itself (Hunston 2002). In order to be significant, statistical tests, normally log-likelihood (LL) and log-ratio (LR) tests, determine their frequency (Hunston 2002; McEnery *et al.* 2006). Since one statistic for collocation extraction may yield skewed results, we applied both LL and LR in order to obtain more reliable data.

3.3.1. General data across the BNC

In order to gain an overview of the use of the three lexical items in the entire BNC as well as across specific genres, we applied both AF and RF, dispersion measures (mean frequency, standard deviation and Julliand's D) as well as a simple text count (Brezina *et al.* 2015). We also investigated the ratio between singular and plural forms of the three lexical items in the whole corpus. In all other studies lemma-based queries were used.

For easier handling of the data, we introduced text categories that differ from the original BNC text types and include the following: ACADEMIC, FICTION, NON-ACADEMIC, NEWS, OTHER WRITTEN and SPOKEN.

3.3.2. Collocation analysis methods

It was assumed that collocations at the textual level have the potential to reveal the semantic features of lexical items and their underlying concepts. Therefore, in order to determine the semantic profiles of DISEASE, ILLNESS and SICKNESS, we conducted six collocation queries, two for each term.

We employed a lemma-based noun-only collocation search, with the collocation window defined as 5L-5R and minimum collocate and node-collocate frequencies set to 5. We used two statistical measures to extract collocations, namely, log likelihood (or LL, employing significance statistics) and log ratio (or LR, measuring the effect size) (see e.g. Evert and Hardie 2014; McInnes 2004). While LL scores collocations by statistical significance, LR measures how big the difference is between the (relative) frequency of the collocate alongside the node, and its (relative) frequency in the rest of the corpus or sub-corpus. We therefore sorted collocates by LL and LR: the presence of a collocate in both lists was the inclusion criterion for

collocates in the present study. For the purpose of the semantic analysis, statistical significance was set as $p \leq 0.05$.

The next step after generating collocation lists for each statistical measure was to compare the results and extract the top 150 common collocates from both lists. The collocates were then analysed for semantic preference, taking a corpus-driven approach. Altogether, 23 semantic categories were identified in the study (see Table 1) and each collocate was assigned to one of them. The list was extended every time a new category emerged from the analysis. The final step included calculating the most frequent semantic categories for each term in order to draft their semantic profiles.

Semantic tag	**SEMANTIC CATEGORY**
TRANSMISSION	how it develops, contagious, how it is transmitted, inherited
NAME	specific name of a disease/illness/sickness
BODY PART	reference to a body part, location of the disease/illness/sickness in the body function of the body
EFFECT	reference to symptom, effect of the disease/illness/sickness
CAUSE	reference to cause of the disease/illness/sickness
LACK	reference to a lack of food
QUALITY	quality or characteristics of the disease/illness/sickness
TYPE	reference to the type of disease/illness/sickness
WHO/WHAT	reference to whom or what is ill
FUNCT. WD.	function word
GEOGRAPHY	geography
QUANTITY	quantity
TREATMENT	treatment
LEGAL	reference to a legal document, legislation
TIME	reference to age, time period in life
PREVENTION	prevention
SYNONYM (SG)	sg that is similar to disease/illness/sickness
DIFFERENT	different, various, other
EXAMPLE	example, such as, including
SPECIFICITY	specific, certain
INCIDENCE	incidence, case, one specific example
RISK	risk
ONSET	start, onset of a disease
INSTITUTION	institution
SUPERSTITION	popular belief, superstition

Table 1
Semantic categories identified for the most frequent collocates
of DISEASE, ILLNESS and SICKNESS in the BNC.

Having explained our methodological approach, we will now turn to the results of the data analysis in the next paragraphs.

4. Results and Discussion

4.1 General comparison of the collocational behaviour of the three lexical items

Table 2 below presents the overall frequencies of the three lexical items analysed in this study. Please consider that the different frequency results from those indicated in paragraph 3.2.1. are due to the fact that the BNC BYU is a different version of the BNC CQP Web platform, one which seems more complete and offers a wider range of options and apparently more reliable results (cf. also https://www.uni-bamberg.de/fileadmin/eng-ling/fs/Chapter_11/Index.html?3123ExerciseforBYUBNC.html [09/12/18]).

As can be seen from the summary in Table 2, there seems to be a strong preference for the use of *disease* rather than *illness* and *sickness*, and overall, the relative frequency is higher in the written sub-corpus than in the spoken one. There is also a marked preference for the use of lexical items in the singular form.

Lemma	No. of texts	Absolute Frequency (spoken)	Relative Frequency/million words (written/spoken)	singular/plural (%)
disease	1,214	10,680 (291)	95.27 (103.8/24.3)	83 / 17
illness	1,029	3,718 (214)	33.17 (35/17.9)	86 / 14
sickness	528	1,209 (101)	10.78 (11.7/8.4)	99 / 1

Table 2
Absolute and relative frequencies of *disease*, *illness* and *sickness* in the BNC.

As regards the types of texts under consideration, details are given in Table 3 below. Each text type, i.e. ACADEMIC, FICTION, NON-ACADEMIC, NEWS and other written texts, together with SPOKEN ones, has been investigated in relation to the absolute frequencies of the three lemmas.

Lemma	Academic (RF)	Fiction (RF)	Non-academic (RF)	News (RF)	Other written (RF)	Spoken (RF)
disease	4,994 (281.2)	328 (17)	2,591 (95.2)	645 (60.9)	1,831 (72.7)	291 (24.3)
illness	940 (52.9)	310 (16)	1,170 (43)	382 (36)	702 (27.9)	214 (17.9)
sickness	279 (15.7)	157 (8.1)	286 (10.5)	98 (9.25)	288 (11.4)	101 (8.4)

Table 3.
Absolute and relative frequencies across text-types

As can be seen in Table 3 above, *disease* seems to be the most frequently used lemma throughout all text types constituting the BNC. However, while *disease* has a higher frequency in ACADEMIC, NON-ACADEMIC and

miscellaneous written texts, the highest frequency of *illness* is to be found in ACADEMIC texts, followed by NON-ACADEMIC and NEWS text types. *Sickness*, on the other hand, is mainly found in ACADEMIC, miscellaneous and NEWS WRITTEN texts. This seems to suggest that there is a difference in use and that the three lemmas can be regarded as synonyms only in particular text types. The spoken sub-corpus confirms the top presence of *disease*, followed by *illness* and *sickness*.

Table 4 below presents the statistical tests we carried out to measure dispersion and distribution across text categories.

	[DISEASE]	[ILLNESS]	[SICKNESS]
Mean frequency	79.02	30.2	11.66
Standard deviation	81.23	14.22	6.76
Juilland's D test	0.61	0.82	0.78

Table 4
Dispersion across text categories

Based on mean frequency, the most frequent word is again *disease*; however, it is not as evenly distributed across the text categories investigated. Juilliand's D for *disease* is the lowest due to academic text bias.

In the following paragraphs, the results of the collocation analysis of the three lexical items will be discussed.

4.2 Semantic Profiles

4.2.1 Disease

Overall, there were 10,680 occurrences of the lemma *disease*, of which 8,855 were used in singular form and 1,825 in plural form. As aforementioned, the analysis of the lemma and both word forms of *disease* started out with the first 150 collocates identified by the two statistical measures LL and LR. Collocates appearing in both lists were selected for semantic analysis. This reduced list contained 89 collocates in the case of the lemma, 93 in the case of the singular form and 135 collocates in the plural form. These results reveal that although the number of occurrences of the plural form was about a quarter of the number of occurrences of the singular form, the plural form seems to be more productive in terms of the number of different collocates.

The qualitative analysis of the collocational patterns of *disease* in terms of the semantic preference of its different forms yielded interesting results. As can be seen in Table 5, altogether, 23 different semantic categories were identified among the collocates of all forms of *disease*.

SEMTAG	Semantic category	Number of collocates (*disease*)	Number of collocates (*diseases*)	Examples
BODY PART	reference to body part, location of the disease in the body	31	11	*arterial, bladder, gall, gastrointestinal*
NAME	specific name of a disease	17	10	*Alzheimer, Creutzfeldt-Jakob, diabetes, legionnaire, HIV, malaria, measles*
QUALITY	quality of the disease	15	20	*addictive, autoimmune, malignant, severity, acute*
TRANSMISSION	how the disease develops or is transmitted	9	20	*communicable, inherited, transmitted, blood-borne, insect-borne, infectious*
EFFECT	reference to symptom, effect of the disease	8	10	*obstructive, suffering, symptoms, ulcerative, die*
CAUSE	reference to cause of the disease	4	8	*alcoholic, pathogenesis, accidents, cause caused, viral*
WHO/WHAT	reference to whom or what is ill	2	8	*sufferer, elm, animals, cattle, fish, plant*
FUNCT. WD	function words	0	10	*against, among, from, which*
TREATMENT	treatment	2	8	*diagnosis, clinics, treat, treatment*
QUANTITY	quantity	1	8	*sporadic, prevalence, rare*
PREVENTION	prevention	0	5	*combat, drugs, prescribed, prevent*
SYNONYM	sg that is similar to disease	1	3	*pests, illnesses*
DIFFERENT	different	0	3	*different, various, other*
EXAMPLE	example	0	2	*such as, including*
SPECIFICITY	specific	0	2	*specific, certain*
GEOGRAPHY	geography	0	2	*tropical, Western*
LACK	reference to lack of food	1	1	*malnutrition, starvation*
LEGAL	reference to legal documents	0	1	*acts*
TIME	reference to age	0	1	*childhood*
OTHER	one specific case	0	1	*incidence*
RISK	risk	0	1	*risk*
ONSET	start a disease	1	0	*onset*
INSTITUTION	Institution	1	0	*centre*

Table 5
Semantic profile of *disease* based on collocational analysis.

The highest number of categories was identified among collocates of the plural form, which directly corresponds to the highest number of collocates for *disease*. As can be seen from the data, most of the collocates of the plural form are associated with the quality and characteristics of the disease and how it develops or is transmitted. The singular form, however, is most frequently associated with a body part affected by the disease. Collocates of the singular form are often the specific names of diseases, for example,

Alzheimer, Creutzfeldt-Jakob and *malaria*. The semantic categories of symptoms or effects, represented by such collocates as *affecting, crippling, deaths* and to a lesser extent the causes of the disease, illustrated by collocates such as *causes, fungal, parasitic, smoking, viral,* are equally represented among the collocates of the singular and plural forms. Interestingly, a few function words, such as *and, are, as, from, of, these*, collocate with the plural rather than the singular form of *disease*. Fewer collocates are related to treatment, for example, *cure, diagnosis, hospital, treat, treating*, prevalence (quantity) of the disease, for instance, *common, many, multiple, number* and who or what is ill, examples of which include *animals, cattle, fish, horses, human* and *patients*.

4.2.2 Illness

Overall, there are 3,718 occurrences of the lemma *illness,* of which 3,208 are used in the singular form and 510 in the plural form. Based on a comparison of the numbers of collocates generated by the two statistical measures (LL and LR), the first 136 collocates for singular form and 47 collocates for plural form were included in the analysis. The collocates were classified into 11 semantic categories. The relevant semantic categories with a few examples are presented in Table 6, below. The semantic categories include, for example, specific names of illnesses, such as *schizophrenia*, *asthma* or quality, characteristics of an illness, such as *dangerous illness, serious illness, common illness*, as well as the cause of the illness, such as *injury* or *HIV*-related illnesses.

The collocation analysis of *illness* has revealed that there are no significant differences between the singular and plural forms. Mostly, *illness* is used in relation to psychological and mental illnesses, for example, *mental illness, psychosomatic illness*, *depressive illness*. In addition, the collocates of *illness* also describe the symptoms and effects of such illnesses, for example: *life-threatening illness, depressive illness, long-term illness.* Another important semantic category of the collocates of the semantic profile of *illness* is about dealing with *illness(s),* for example, *treat illness(s), recover from illness(s), prevent illness, overcome illness, cope* with *illness, diagnose illness.*

SEMTAG	**Semantic category**	**Number of collocates (*illness*)**	**Number of collocates (*illnesses*)**	***Examples***
FUNCT. WD	Function word	34	19	*or, often, with*
EFFECT	Reference to symptom and/or effect of the illness	28	5	*life-threatening, flu-like*
DEAL	Dealing with illnesses	22	4	*treat, recover, diagnose*
TYPE	Type of illness in general	16	5	*mental, recurring*
QUALITY	Quality, characteristics of illness	15	4	*dangerous, serious*
CAUSE	Reference to cause	11	4	*injury, HIV-related*
WHO/WHAT	Reference to people	9	3	*patient, childhood, family*
NAME	Specific names of illnesses	5	1	*schizophrenia, asthma*
BODY PART	Reference to body part, location of the disease in the body, function of the body.	4	1	*brain, respiratory*
TRANSMISSION	How illness develops, transmits etc.	4	1	*infectious, enteric*
LACK	Reference to lack of resources (usually food)	1	0	*poverty*
INSTITUTION	Institution	1	0	*hospital*

Table 6
Semantic profile of *illness* based on collocation analysis.

4.2.3 Sickness

Overall, the BNC indicates 1,209 occurrences of the lemma *sickness*, of which 1,205 are used in the singular form and only 14 occurrences in the plural form.

The procedure for the semantic analysis of *sickness* was identical to that adopted for the analyses of *disease* and *illness*. As for the previous lemmas taken into consideration, collocates were detected with both LL and LR; this resulted in a reduced list of 79 collocates, of which 77 can be found in the case of the singular form *sickness*, and 4 in the case of the plural one. All the occurrences have been grouped into 10 semantic categories, as can be seen in Table 7, below.

SEMTAG	Semantic Category	Number of collocates (*sickness*)	Number of collocates (*sicknesses*)	*Examples*
LEGAL	Reference to sickness in legal terms: job and sickness allowance, benefits, rights, insurance, social security etc.	36	0	*absence, absenteeism, allowance, invalidity, rates (of sickness absence), statutory sick pay*
TYPE	Reference to type of sickness	15	1	*altitude, decompression, radiation, spells*
FUNCT. WD	Function words: preposition/conjunction	8	1	*among, and, for, from, of, overall, through*
CONDITION	Human condition	5	1	*age, death, health, ill,* also metaphorical: *the State's sicknesses*
DEGREE	Degree of sickness	5	0	*bout, levels, days, grade*
EFFECT	Reference to symptom, effect	3	0	*diarrhoea, effects, symptoms*
QUALITY	Quality of the disease, characteristic of a disease	3	0	*chronic, acute, long*
SYNONYM	Synonym	1	0	*illness*
SUPERSITION (in literary contexts)	Popular belief, superstition	0	1	*evil*

Table 7
Semantic profile of *sickness* based on collocation analysis.

The data suggest that the collocates of the singular form are associated with a wider range of semantic sets. The most prevalent use is linked to those types of *sickness* which may affect professional life, for example, *allowance*, *benefits*, *insurance*. This seems to indicate that sickness is often used to refer to a state of health in a legal sense. Even when the semantic profile refers to the type and degree of sickness, both its symptoms and characteristics are related to the types of sickness that affect employment life from an insurance or pension-system point of view.

The plural form of *sickness* co-occurred with only four different collocates. These refer to a type of condition that has led to sickness and is related to superstition. It must be said, however, that the plural forms occur in a spoken classroom context in which people are commenting on a literary text. However, given the extremely low frequency of *sickness* in the plural form, far-reaching generalizations cannot be drawn.

Overall, our findings reveal that there are considerable differences between the frequencies, numbers of collocates and which text types the selected near synonymous lexical items are frequently used in. In addition, it was found that the collocational patterns of the examined lexical items also show marked differences in the numbers of collocates and their semantic preferences. This corresponds to earlier studies that suggest that near

synonyms exhibit different collocational behaviours and semantic preferences (Xiao and McEnery 2006). Previous research on the collocational patterns of lexical items suggests that individual word forms of the node word often collocate with different words (Hoey 2005; Gledhill 2000; Sinclair 1991; Tognini-Bonelli 2001). The health-related words examined show similar collocational behaviours, as their singular and plural forms exhibit different collocational behaviours in terms of both collocates and semantic preference.

5. Conclusions

The study has revealed that the three terms under investigation, despite being seen as near synonyms, differ in their collocational behaviours and therefore exhibit different semantic preferences. Overall, *disease* was found to be the most frequent of the three terms. Several semantic categories were identified among its collocates and there is a marked difference in the number of semantic categories associated with the plural and singular forms, the plural form being more productive in terms of both the number of individual collocates and semantic categories. These categories indicate that the plural form shows a semantic preference for how diseases are spread and what they are like. At the same time, the singular form has a semantic preference for the semantic category of body parts and the names of types of disease, as for instance indicated in excerpts (1), (2) and (3) taken from the BNC (emphasis in the original texts):

(1) Not only are you much more likely to die from **lung cancer** or **heart disease**, but other illnesses highlighted in this booklet, including cervical cancer, are associated with smoking (A0J_1708)

(2) Now one of the **auto-immune diseases** that has been recognised is erm unusual baldness — it's called alopecia (KRF_662)

(3) Er, this particular **disease**, **Alzheimer's** disease was identified by Jim, who was the deputy mayor, a member of ours, who spoke to you earlier in the week […] (KM0_688)

The most prevalent use of *illness* is related to psychological/mental illnesses and, in particular, to their symptoms and effects. The collocation analysis of *illness* revealed that words that it co-occurs with are more about dealing with *illness(s),* unlike *disease* and *sickness*, as excerpts (4)-(7) seem to suggest:

(4) Officers had undertaken a review of the policies in both mental **illness** and **mental** handicap in response to the 1975 White Paper and 1976 priority services recommendations. (CS7_1027)

(5) The Royal Commission on Mental **Illness** and **Mental** Deficiency introduced the concept of guardianship, and the Mental Health Act 1959 gave the guardian wide powers of control. (EA1_151)

(6) Caring for a relative with a progressive, relapsing **illness** or **terminal** condition makes future related adverse events almost certain (J14_1281)

(7) It is beyond dispute that advances in medicine and improvements in living conditions have enabled individuals who at previous times would not have survived severe **illness** or **chronic** handicaps to live on, perhaps with some disability, into their seventh, eighth, and ninth decades. (CK_187)

These categories indicate that the singular form of *illness* has a semantic preference related to society, as it indicates how a disease may affect professional life and how this has to be regarded within pension-system or insurance-benefit contexts, as can be seen in example (8), below:

(8) If you are disabled by **illness** or **injury** at the time that you enter the agreement, cover will not begin until you return to full time work. (AYP_2356)

As far as *sickness* is concerned, the semantic categories identified among its collocates show that the term is mainly used in the singular form, which is the only productive one:

(9) AIDS touches areas of **sickness**, death and personal behaviour. (A01_513)

(10) **Sickness**, diarrhoea and some drugs may stop it working, and extra precautions must be used. (A0J_366)

(11) If you are entitled to a **sickness allowance** under the occupational sick pay scheme, SSP is paid as part of that sickness allowance. (HD2_2055)

The plural form is found in the spoken corpus only and is linked to a comment made in relation to a textual analysis:

(12) Erm, but in fact, she's she's missed the third sentence and, where she said that the rose has withstood many **sicknesses** and **evils**, erm, whereas in fact, what it says is it withstands and succours against sicknesses and evils, which is a totally different element.

Strangely enough, the collocates in the BNC seems to tell a different story for *sickness* from those indicated by the dictionaries we consulted: sickness has apparently less to do with a temporary condition and more with a socio-economic one.

Our study seems to bring to light fine-grained differences in the meanings of lexical items, making it possible to achieve a far higher level of precision of a sense of disambiguation, for example, in reference works, such as dictionaries, and better matching in an interlingual quest for equivalence.

The method of semantic profiles, as outlined in this paper, has proved to be especially effective for the task at hand. As we have seen in the excerpts above, *disease* is in most cases accompanied by its scientific or popular name, and is mainly used in relation with the body parts affected by it, also to indicate its characteristics and (particularly in the plural form) how it is transmitted. By contrast, the term *illness* indicates the type of disease, its health effects and how these have to be treated. *Sickness* is preferred when the speaker wants to show the effect the disease has on professional life (allowance, benefits, insurance etc.), and therefore has more social implications.

Although this study has some limitations, primarily the small number of lexical items examined, it nevertheless offers interesting insights into how semantic profiles can be outlined. This can be helpful for research in translation studies and language teaching, as it grants lexical and semantic completeness for the terms under investigation.

As a next step, knowing that the BNC is just *one* of the available resources we have at our disposal and acknowledging that the lingua franca of medicine is English, in both its British and American varieties, we aim to compare the findings resulting from a similar investigation to be carried out on COCA to see the extent to which, if any, (a) meanings can vary and (b) whether meaning variations associated with these items depend on the interactants (i.e. professionals/laymen).

Acknowledgements: We would like to thank the anonymous reviewers for their helpful and constructive comments that greatly contributed to improving the final version of the paper. We are also grateful to the editors for their support, generous help and impressive feedback. We would also like to thank Vaclav Brezina, from Lancaster University, for his great support when we decided to write this paper.

Bionotes:
Stefania M. Maci is Associate Professor of English and pro Vice Chancellor for Education at the University of Bergamo. Her main research focusses on medical and tourism discourse, Critical Discourse Analysis, Corpus Linguistics, Discourse Analysis, Multimodality.

Dr Reka R. Jablonkai works at the University of Bath. She is Director of Studies of the MA TESOL programme. Her research interests include corpus-based discourse analysis, corpora in English language teaching and learning, ESP and EAP, and intercultural communication. She is grant holder of the BAICE Seed fund.

Marek Łukasik, PhD, assistant professor at the Department of English Studies, Institute of Modern Languages, Pomeranian University in Słupsk, Poland. Head of Laboratory for

Modern Methods in Applied Linguistics. His main research interests include metalexicography, terminography, corpus linguistics and specialised translation.

Sophiko Daraselia is Doctoral Researcher and Teaching Assistant at the University of Leeds, UK. Her research interests include corpus linguistics, lexicography and morphology, with particular focus on corpus design and construction and corpus analysis methods, software tools and application of corpus methods in lexicography.

Daniel Knuchel is a PhD Candidate in German Linguistics at the University of Zurich. His main research interests are Discourse Analysis, Corpus Linguistics and Semantics.

Authors' addresses: stefania.maci@unibg.it; reka.jablonkai@gmail.com; marek.lukasik@apsl.edu.pl; mlsd@leeds.ac.uk; daniel.knuchel@ds.uzh.ch

References

Baethge C. 2008, *The Languages of Medicine*, in "Deutsches Ärzteblatt International" 105 [3], pp. 37-40.

Baker P., Gabrielatos C. and McEnery T. 2013, *Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word 'Muslim' in the British Press 1998-2009*, in "Applied Linguistics" 34 [3], pp. 255-278.

Baker P., Gabrielatos C., Khosravinik M., Krzyżanowski M., McEnery T. and Wodak R. 2008, *A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press*, in "Discourse & Society" 19 [3], pp. 273-306.

Barron A. 2003, *Acquisition in interlanguage pragmatics*, Benjamins, Amsterdam.

Bednarek M. 2008, *Semantic preference and semantic prosody re-examined*, in "Corpus Linguistics and Linguistic Theory" 4 [2], pp. 119-139.

Brezina V., McEnery T. and Wattam S. 2015, *Collocations in context: A new perspective on collocation networks*, in "International Journal of Corpus Linguistics" 20 [2], pp. 139-173.

Busse D. 2015, *Sprachverstehen und Textinterpretation*, Springer VS, Wiesbaden.

Candlin C.N. and Candlin S. 2003, *Health care communication: A problematic site for Applied Linguistics research*, in "Annual Review of Applied Linguistics" 23, pp. 134-154.

Collins English Dictionary. https://www.collinsdictionary.com/dictionary/english (17.3.2018).

Crystal D. 2002, *The Cambridge Encyclopedia of the English Language*, Cambridge University Press, Cambridge.

Davies M. 2004, *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). https://corpus.byu.edu/bnc/ (17.3.2018).

Djulbegovic B. and Guyatt G.H. 2017, *Progress in evidence-based medicine: a quarter century on*, in "Lancet" 390, pp. 415-423.

Evert S. and Hardie A. 2014, *Keyword and collocation statistics: Under the hood of CQPwep*. http://ucrel.lancs.ac.uk/talc2014/doc/Workshop-STATS.pdf (17.3.2018).

Firth J. 1957 [1951], *Modes of meaning*, in Firth J., *Papers in Linguistics, 1934-1951*, Oxford University Press, Oxford, pp. 190-215.

Gablasova D., Brezina, V. and McEnery T. 2017, *Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence*, in "Language Learning" 67 [S1], pp. 155-179.

Gledhill C. 2000, *Collocations in science writing*, Gunter Narr, Tübingen.

Gotti M. 2016, *Variations in Medical Discourse for Academic Purposes*, in Ordóñ-López E. and Edo-Marzá N. (eds.), *Medical Discourse in Professional, Academic and Popular Settings*, Multilingual Matters, Bristol-Buffalo-Toronto, pp. 9-30.

Graddol D. 2006, *English Next. Why global English may mean the end of English as a foreign language*, British Council. www.britishcouncil.org/learning-research-english-next.pdf (17.3.2018).

Gries S. 2008. *Dispersions and adjusted frequencies*, in "International Journal of Corpus Linguistics" 13 [4]: 404-438.

Hardie A. 2012, *CQPweb - combining power, flexibility and usability in a corpus analysis tool*, in "International Journal of Corpus Linguistics" 17 [3], pp. 380-409.

Hoey M. 2005, *Lexical priming*, Routledge, New York.
Hunston S. 1995, *A corpus study of some English verbs of attribution*, in "Functions of Language" 2, pp. 133-158.
Hunston S. 2002, *Corpora in applied linguistics*, Oxford University Press, Oxford.
Loiacono A. 2018, *Expressing Health, Disease and Illness in English*, in Loiacono A., *The Medical Alphabet: An English textbook on healthcare in the digital age*, Matarrese Editore, Andria, pp. 382-405.
McEnery T., Xiao R. and Tono Y. 2006, *Corpus-based language studies*, Routledge, New York.
McInnes B.T. 2004, *Extending the Log Likelihood Measure to Improve Collocation Identification*, University of Minnesota. http://www.d.umn.edu/~tpederse/Pubs/bridget-thesis.pdf (17.3.2018).
Merriam-Webster Dictionary. https://www.merriam-webster.com/ (17.3.2018).
Nelson M. 2006, *Semantic association in Business English: A corpus-based analysis*, in "English for Specific Purposes" 25 [2], pp. 217-234.
Oxford English Dictionary. http://www.oed.com/ (17.3.2018).
Partington A. 2004, *'Utterly content in each other's company' Semantic prosody and semantic preference*, in "International Journal of Corpus Linguistics" 9, pp. 131-156.
Sentel T. and Braun K.L. 2012, *Low health literacy, limited English proficiency, and health status in Asians, Latinos, and other racial/ethnic groups in California*, in "Journal of Health Communication" 17 [3], pp. 82-99.
Sinclair J. 1991, *Corpus, concordance, collocation*, Oxford University Press, Oxford.
Sinclair J. 2004, *The lexical item*, in Sinclair J. (ed.), *Trust the Text. Language, Corpus and Discourse*, Routledge, London and NewYork, pp. 131-148.
Stubbs M. 1995, *Collocations and semantic profiles: On the cause of the trouble with quantitative studies*, in "Functions of Language" 2 [1], pp. 1-33.
Stubbs M. 2001, *Words and Phrases: Corpus Studies of Lexical Semantics*, Blackwell Publishers, Oxford.
Talbott J.H. 1962, *The British Medical Dictionary*, in "JAMA" 180 [11], pp. 990.
Taavitsainen I. 2018, *Meaning-making practices in the history of medical English: A sociopragmatic approach*, in "Journal of historical pragmatics" 18 [2], pp. 252-270.
Tognini-Bonelli E. 2001, *Corpus linguistics at work*, John Benjamins Publishing Company, Amsterdam and Philadelphia.
Xiao R. and McEnery T. 2006, *Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective*, in "Applied Linguistics" 27 [1], pp. 103-129.
Wittgenstein L. 2003 [1953], *Philosophische Untersuchungen*, Suhrkamp, Frankfurt aM..
Wulff H.R. 2004, *The language of medicine*, in "Journal of the Royal Society of Medicine" 97 [4], pp. 187-188.